A Social Robot System for Modeling Children's Word Pronunciation

Socially Interactive Agents Track

Samuel Spaulding, Huili Chen, Safinah Ali, Michael Kulinski, Cynthia Breazeal
{samuelsp, hchen25, safinah}@media.mit.edu, kulinski@mit.edu, cynthiab@media.mit.edu
MIT Media Lab
Cambridge, MA

ABSTRACT

Autonomous educational social robots can be used to help promote literacy skills in young children. Such robots, which emulate the emotive, perceptual, and empathic abilities of human teachers, are capable of replicating some of the benefits of one-on-one tutoring from human teachers, in part by leveraging individual student's behavior and task performance data to infer sophisticated models of their knowledge. These student models are then used to provide personalized educational experiences by, for example, determining the optimal sequencing of curricular material.

In this paper we introduce an integrated system for autonomously analyzing and assessing children's speech and pronunciation in the context of an interactive word game between a social robot and a child. We present a novel game environment and its computational formulation, an integrated pipeline for capturing and analyzing children's speech in real-time, and an autonomous robot that models children's word pronunciation via Gaussian Process Regression (GPR), augmented with an Active Learning protocol that informs the robot's behavior.

We show that the system is capable of autonomously assessing children's pronunciation ability, with ground truth determined by a post-experiment evaluation by human raters. We also compare phoneme- and word-level GPR models and discuss trade-offs of each approach in modeling children's pronunciation. Finally, we describe and analyze a pipeline for automatic analysis of children's speech and pronunciation, including an evaluation of SpeechAce as a tool for future development of autonomous, speech-based language tutors.

KEYWORDS

Social robot; Student Modeling; Human-Robot Interaction; Speech-based Systems; Intelligent Tutoring Systems

ACM Reference Format:
Samuel Spaulding, Huili Chen, Safinah Ali, Michael Kulinski, Cynthia Breazeal. 2018. A Social Robot System for Modeling Children's Word Pronunciation. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 9 pages.

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

1 INTRODUCTION

Developing robots capable of interactive educational play could have profound impact for access and scalability of high-quality early childhood education. Such robots can combine the scale and precision of educational software with the physical embodiment and sensing of good human teachers. In addition, robot learning companions can use sophisticated models of individual student's performance and ability to provide personalized educational experiences, e.g., adapting to an individual's preferred teaching style or introducing new curricular elements at a optimal pace for each student.

Typically, such student models are inferred by directly asking the student to demonstrate their knowledge with test questions, each of which may be linked to a particular piece of knowledge in the curriculum. However, when the curriculum is large, directly testing student knowledge may be infeasible or unpleasant. The design of "digital workbook" style software is well-suited for maximizing the number of testing opportunities, but often leads to disengaged behaviors, colloquially known as "gaming the system" [1]. Students frequently exhibit a high degree of engagement in interactions with socially expressive robots, which has led to significant research into developing social robot tutors to provide more personalized, interactive, and engaging educational experiences. Though it is increasingly feasible to deploy autonomous robot tutors in schools ([2], [15], [10]), it is still challenging for such systems to leverage the real-time, situated nature of the interaction by sensing natural interaction cues. As sensing technologies become more widespread and accessible, integrating this data into student models will lead to more natural and effective digital tutoring agents.

Already, researchers are developing autonomous robot tutors that can analyze facial expressions, electrodermal activity (EDA), or body pose (e.g., [5, 33]). One interaction modality with potential for particularly strong impact is speech. Despite the importance of speech as a channel for natural communication, as well as the rapid increase in performance and adoption of speech-based technologies by adults, modeling children's speech remains a difficult and open research problem.

1.1 An interactive robotic system for collecting and analyzing children's speech

In this paper, we present an autonomous robotic tutoring system designed to evaluate and model children's productive vocabulary by recording and analyzing their spoken responses. It combines **1)** an interactive tablet game interface, **2)** word- and phoneme-level student models based on Gaussian Process Regression (GPR)

for children's word pronunciation, **3)** an active learning protocol for efficiently determining and introducing adaptive, personalized content, **4)** a sensor suite and automated analysis pipeline for capturing and analyzing children's speech and pronunciation based on Amazon Mechanical Turk and the SpeechAce pronunciation software and **5)** a physically-embodied social robot and associated behaviors to facilitate an engaging and entertaining educational interaction. To our knowledge, this is the first work to rigorously and autonomously model children's spoken word pronunciation in the context of an interactive game with a social robot.

The design of the game, built on an Android tablet, is inspired by the principles of peer-to-peer learning and educational play. Thus, the interaction is framed as a competitive game between a peer-like robot (an expressive robotic 'character' with a child-like voice and expressive sound effects, see Sec: 3.1) and the child. The competitive design of the game and its variable reward structure help mask the sometimes tedious nature of collecting data for student models.

To evaluate whether the system is capable of sustaining a speech-based interaction with children and accurately modeling their pronunciation, we conducted an experiment with 15 children, in which each child played several rounds of an interactive word game called "WordRacer" on an Android tablet against a robot (Fig. 1). The child and the robot sit across from each other with the tablet in between and a "ring-in" button on each side. Pictures and letters of words from a Test Curriculum (e.g, animals or common household items, see Sec. 3.3) appear in the center of the tablet, and when a picture/word pair appears, the robot and child "race" to tap their button and "ring in", giving the player who rings in first a chance to say the word out loud and receive a point if the pronunciation is deemed "correct". The design of the game allows us to collect and analyze children's speech samples in a natural and engaging fashion during gameplay.

The robot analyzes the collected speech from each child's in-game responses to construct two different models of word pronunciation ability: a PHONEME-GPR model, which models the child's average pronunciation score of each *phoneme* in the standard ARPAbet list of English phonemes as a normal random variable, and a WORD-GPR model, which models the average pronunciation score of each *word* in the Test Curriculum as a normal random variable. Note that while a PHONEME-GPR model can be transformed into a WORD-GPR model (by averaging the posterior prediction mean of each phoneme in a word), the structure and training of each model is different.

In addition to modeling children's pronunciation, the GPR models provide a mechanism for personalized content sequencing. After each round of the game, the tutoring agent selects the next word via an active learning protocol, based on one of the GPR models. Active Learning is well-suited for domains in which labeled data is rare, but in which the algorithm can can access unlabeled examples and selectively prompt the user to label a given example. In this case, our unlabeled examples are either un-answered words from the Test Curriculum or ARPAbet phonemes (depending on the model) which the agent has not observed the user respond to. After each round, the word used in the next round is selected from the Test Curriculum by an active learning protocol based on either the WORD-GPR model or the PHONEME-GPR model. During data collection, half of the children played against a robot that used

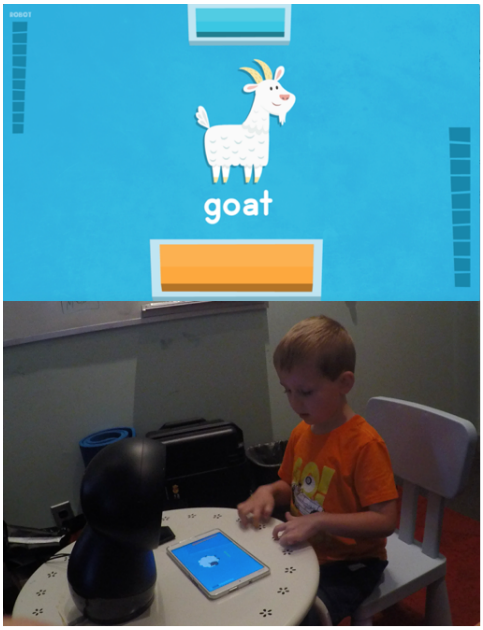

Figure 1: Picture of the interaction setup from which the data was collected and system was evaluated. Children played a competitive word game with a social robot, followed by a self-paced quiz using only a tablet

the word model (WORD-GPR) to adaptively select the words in the game, while the other half played against a robot that used the phoneme model (PHONEME-GPR) to select words. We refer to these conditions as the Word-Led and Phoneme-Led conditions, respectively.

After the experiment, each child completed a comprehensive post-test evaluation in which the child was prompted to pronounce each of the words in the Test Curriculum (45 words in total) in sequence. The recorded audios were then uploaded to Amazon Mechanical Turk and the pronunciation of the intended word was rated from 0-5 by human experts. This comprehensive examination serves as the "ground truth" by which we evaluate our model's post-game accuracy. We show that SpeechAce provides an acceptable off-the-shelf classification solution at the phoneme level for children's speech, and that student models based on Gaussian Process Regression can quickly and effectively assess student's productive vocabulary across a large curriculum from speech samples recorded and analyzed during an interactive game. Sec. 4.1 contains more detail on the evaluation procedure and analysis pipeline.

With this project, we advance research into alternate methods of assessing child vocabulary from speech in the context of an engaging, interactive game. To further the impact of our work, we are releasing the source code and digital assets for the game, as well as an integrated suite of software to support rapid implementation, integration, and evaluation of additional Student and Agent models for assessing, modeling, and responding to children's speech. We hope the tools we have developed will advance the broader goals of developing autonomous agents capable of facilitating speech-based educational interactions with children.

2 RELATED WORK

Here we review work related to analysis of children's productive vocabulary, computational inference of student models, and other adaptive educational child-robot interaction systems.

2.1 Modeling children's vocabulary from speech

Learning to read is one of the most important tasks of early childhood education, but in 2013, only 35% of 4th, 36% of 8th and 38% of 12th grade students in the USA achieved proficiency in reading on the National Assessment of Educational Progress (NAEP) [23]. Pre-literacy skills, like phonological awareness, alphabetic and vocabulary knowledge, are important precursors to the development of literacy skills in later grades and can help prevent academic failure [4, 8, 14, 24, 32]. Social robot learning companions have been proposed as a technological intervention to help address this gap [34], based on the ability of such systems to personalize and provide adaptive content as well as the scale and ease-of-use afforded by social technologies. In order to fulfill the promise of long-term, personalized tutoring interactions, such robots must possess techniques for engaging, age-appropriate, and autonomous assessment [37]. Our system is designed to address the problem of rapidly assessing children's vocabulary and pronunciation across a large curriculum of age-appropriate words.

Assessing children's vocabulary and pronunciation is important but difficult. Comprehensive tests such as the Peabody Picture Vocabulary Test (PPVT) are among the most widely used methods to asses a child's vocabulary level. The PPVT requires asking children to identify depictions of increasingly difficult target words until they fail to correctly identify a fixed percentage of words; a full administration can require prompting a child for an answer over 200 times. Moreover, because children indicate their answer by pointing rather than speaking, the PPVT measures only children's 'receptive' vocabulary. Assessments of 'productive' vocabulary — the words that a child can speak aloud and pronounce correctly — are often conducted in a clinical setting by a trained speech therapist.

Compared to adults, children's speech is far more idiosyncratic and less regular, featuring frequent disfluencies (such as 'Um' or 'Uh'), omissions, mispronunciations, and other irregular language patterns. Physiologically, children are more diverse in their motor development and exhibit greater variation in spectral and temporal parameters [21]. A recent analysis by Kennedy et al. concluded that ASR systems which use language models trained on adults do not perform well in similar interactions with children and, consequently, speech-based interactions between children and digital agents are more difficult to implement and evaluate [19].

Despite these challenges, progress has been made in modeling and analyzing children's speech. Due to the challenges of maintaining an engaging interaction while simultaneously collecting enough speech data to train a good model, many of the most successful models are derived from specialized feature models collected in non-interactive settings (e.g., [35]). In another recent analysis, Yeung et al. used a template-based model derived from sparse data for analyzing rhotic pronunciations [38]. Fringi et al. analyzed children's phonological errors and determined that while correcting for children's pronunciation errors can improve recognition performance, there remain myriad other sources of model error. Thus, analyzing children's speech remains an open research topic, with a unique set of challenges compared to adult speech [3].

2.2 Autonomous speech-based robot tutors

Developing autonomous literacy tutors has been a long-standing motivation for research on interactive systems that can analyze children's speech [12, 13]. This application is enjoying renewed effort and attention from the research community, as enabling technologies such as robust robot platforms, cloud-based speech interfaces, and algorithms and infrastructure for learning from large datasets develop into mature fields.

Sadoughi et al. created a robot that uses children's speech features to modulate its own vocal behavior during an interactive game with children. However, in line with the recommendations outlined by Kennedy et al. [19], their implementation relied on prosodic features of speech that can be reliably detected, not on semantic or phonetic features, which are more challenging to identify [30]. Gordon & Breazeal used a Bayesian active learning algorithm to model a child's vocabulary and select new words to introduce to a child during an interactive literacy game. However, the interaction did not analyze or rely on children's speech, and outside of the choice of words introduced, the robot's game behavior was scripted [9]. Park et al. developed a robot that learned parameters for a rule-based model of "social backchanneling" from children's speech and gaze cues and evaluated it in the context of a subsequent speech-based interaction between a child and a robot [25].

From this prior research, among others, we highlight two themes. First, situated interactions provide necessary context to help ease the burden of collecting natural, realistic speech data from children. Second, children's speech poses unique modeling challenges which must be overcome to facilitate impactful speech-based tutoring interactions for literacy. Notwithstanding the promising efforts of recent years, the problem of analyzing children's speech in real time during an autonomous interaction remains a challenge for future development of speech-based tutoring systems.

3 SYSTEM ARCHITECTURE OVERVIEW

In this section we give an overview of our integrated system for collecting, analyzing, and modeling children's spoken word pronunciation during interactive gameplay. The full system architecture is outlined in Figure 2.

The system is based on a Model-View-Controller design, with the core system built on top of the open-source Robot Operating System (ROS) framework [28]. The WordRacer **Controller** manages the state of the game, receives input from the **Sensors**, sends data and queries to the Agent and Student **Models** and sends commands to the **View**, the tablet interface itself and the robot hardware, both of which run native software that receive commands to load content and play appropriate visualizations, animations, and sound effects.

The source code of the system and instructions for interfacing with the game can be found at https://github.com/mitmedialab/speech-tapgame-aamas18. All modules communicate via ROS, except for the **View** components (the tablet and robot hardware), which communicate with the Game Controller via ROSbridge, a websocket-based protocol that allows a wider range of web-connected devices to interface with a ROS system [6]. Unlike the broader

agents community, there are few, if any, widely-accepted baseline implementations of common algorithms and modeling techniques for educational tutoring agents and student models, as the educational tasks and environments in which they are evaluated differ considerably across projects. The computational formulation of WordRacer is essentially that of a multi-objective sequential decision-making problem, and its design and architecture are fully extensible for new content, sensors, and modeling techniques. By releasing our system code and assets, we hope to take a step towards fostering reproducibility and shared baselines in the field, and hope that WordRacer and its associated software architecture may serve as a testbed for a variety of models of interactive educational agent behavior and student knowledge.

3.1 WordRacer: task and model overview

In this section we introduce WordRacer, an interactive vocabulary game designed for children age 4-6, and its computational formalism.

Fig. 1 shows a closeup of the game interface and the experimental setup used for evaluation. A robot and child sit across from each other with a tablet in between them. Each side of the tablet has a button, which the players tap to 'ring in' (the robot presses the button digitally, lights up an on-board LED ring, and makes a quick, expressive motion to signal its ring) when an image of a vocabulary word appears in the center of the tablet screen. The first player to tap their button is given an opportunity to say the word. If the word is pronounced correctly, that player 'wins' the round and scores a point, otherwise neither player 'wins' and the game moves on the the next round. The game ends after a fixed number of rounds, and whichever player has won more rounds is declared the winner.

Despite the conceit of the game, the robot's decision to ring in or not, its expressive behavior, and the word introduced in each round are carefully structured by the ongoing inference of the student model and the corresponding active learning procedure.

3.1.1 Agent Model and task formulation. For each round, the robot can choose to ring in and say the word correctly (denoted here as action RING-IN), in essence *demonstrating* to the child how to pronounce the word, or else not ring in (denoted here as action DONT-RING), and let the child demonstrate their ability to pronounce the word. The two actions correspond to different objectives: the RING-IN action helps to teach the child new words by demonstrating a correct pronunciation of a word, while the DONT-RING action informs the agent about the child's state of knowledge by giving the student an opportunity to pronounce words (and subsequently letting the agent assess the pronunciation). The robot chooses which words to teach based on the predictions of the student model, thus refining the student model ultimately helps achieve the teaching objective.

In this initial evaluation, the agent samples from these two actions according to a weighted random sampling method, choosing action DONT-RING with a decreasing weight, ϵ. For this study, we used an initial value of $\epsilon = .8$, with $\Delta\epsilon = -.03$. Thus, in the first round of the game, the robot chooses action RING-IN with probability .2, and in the second round, chooses RING-IN with probability .23. This implementation choice reflects the fact that in the initial rounds of the game, the agent has not yet received enough information to infer an accurate student model, and thus should take actions that allow it to gather information; in later rounds, when the model is more certain, the agent is more likely to ring in and demonstrate word pronunciations. The stochastic nature of the action selection procedure helps maintain the variable reward structure of the game.

From an agent behavior perspective, the game is essentially a multi-objective sequential decision making task: at each round the robot decides what game action to take for that round and which word to introduce to maximize some objective (e.g., gaining information, teaching words, or simply facilitating an engaging interaction). In line with the focus of this paper (evaluating whether the system can effectively assess children's word pronunciation), we use the sampling model described above; future work will address learning an action policy for the game, based on both real-time interaction features and features from the student model.

3.1.2 Active Learning protocol. After the agent has selected an action (RING-IN or DONT-RING), it uses an active learning protocol to select a word from the Test Curriculum best suited to the particular objective. Active learning has been shown to improve the efficiency of learning, especially in machine learning systems in which labeled data is difficult or expensive to obtain. Active learning is also well-suited to domains in which the system designers can computationally characterize the structure of the data space, thereby enabling the active learning system to intelligently select unlabeled points in the input space for querying. Student modeling systems for literacy skills fit both criteria: excessively prompting students to demonstrate their reading/pronunciation ability is tiresome and tedious, and advances in natural language processing have yielded rich models of human language that accurately model many aspects of English phonetics and semantics.

The active learning protocol depends on two variables: which model is driving the active learning (i.e., Phoneme-Led or Word-Led condition) and which action is chosen for the current round (determined by the weighted random sampling method described in Sec. 3.1.1). If the chosen action is RING-IN , then the protocol selects the word from the Test Curriculum with the lowest posterior mean (i.e., the word that the model predicts the child is least likely to pronounce correctly). If the chosen action is DONT-RING , then the protocol selects the word from the Test Curriculum with the highest posterior variance (i.e., the word about which the model is most uncertain). If the WORD-GPR model is driving the active learning, this is a straightforward search over words in the Test Curriculum . In the Phoneme-Led condition, the mean and variance of each word in the Test Curriculum is determined by phonemizing each word and averaging the posterior means or variances in the PHONEME-GPR of the constituent phonemes.

3.2 Pronunciation analysis module

We analyzed children's speech using SpeechAce, a commercial pronunciation analysis software for second-language learners [17]. During gameplay, speech samples are analyzed via the SpeechAce REST API by sending a sample of recorded audio and a phonemized representation of the expected word. SpeechAce computes a set of scores (from [0,100]) representing how well each phoneme was

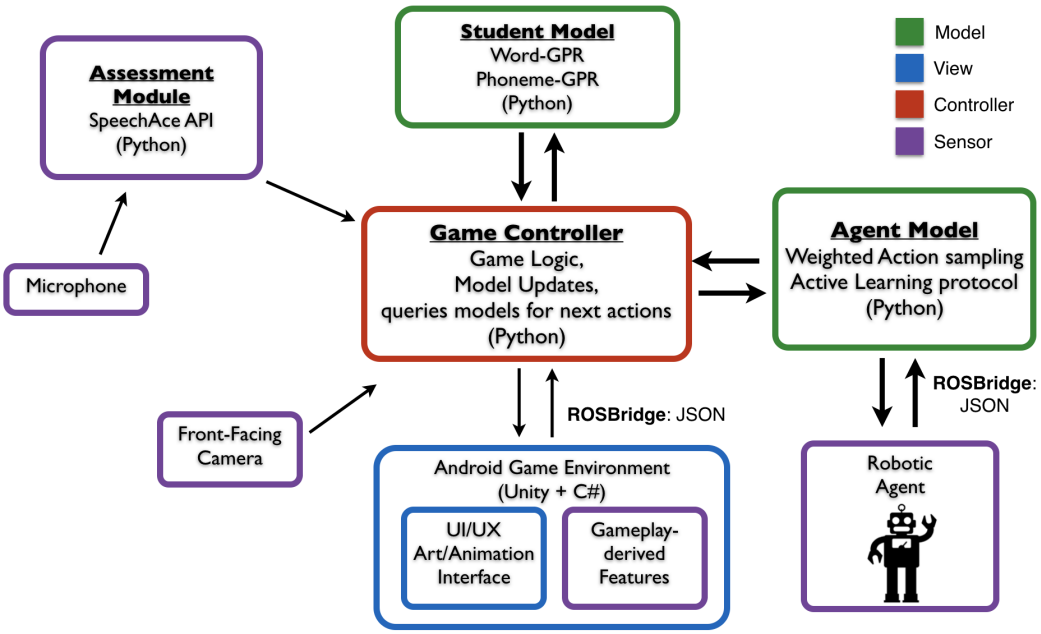

Figure 2: Diagram of interactive system for children's speech analysis. Student and Agent modules are easily reconfigurable to support a variety of models; sensor modules can be easily added or removed. All modules communicate with each other via ROS, except where otherwise noted

pronounced in the word. For example, to analyze a child's pronunciation of the word **'CAT'**, we send the recorded audio file and the phonemization, **["K", "AE", "T"]** to SpeechAce. Sample scores are returned as formatted JSON. All phonemizations are computed within our system, using the ARPAbet phoneme list and the CMU pronouncing dictionary [36]. Each analyzed phoneme and its score is added to the training set of the PHONEME-GPR model, we average the score over all phonemes in the word to compute the score for the WORD-GPR model. As part of our model evaluation, we also compare the results of SpeechAce analysis to the ratings of native English speakers collected via Amazon Mechanical Turk (see Sec. 4.1), to characterize SpeechAce's utility as a tool for further research on autonomous assessment of student word pronunciation.

3.3 Test Curriculum

While the GPR models are extensible to any words for which we can compute pairwise covariance, in this work we evaluate the models on a subset of 45 words for which we have prepared game assets (graphics, sound recordings, and animations). These words constitute both the input domain of training data for the WORD-GPR model and the test domain for both models. For the PHONEME-GPR model, the input domain is the set of 39 ARPAbet phonemes, whose scores are averaged to yield a word model for comparison to human pronunciation rating.

Our motivation for this research is to further the science of intelligent, autonomous, speech-based tutors, with a specific project focus on helping promote early literacy skills in children. Therefore, we chose the initial words in the Test Curriculum in consultation with experts on early childhood literacy and reading, based on pedagogical best-practices. We reviewed typical sight word and word-frequency lists for children in our target age-range, then augmented the list with additional categorical words important for basic conceptual development (e.g., animals, household objects, nature words). With a few exceptions for important 'sight words', we selected words that are highly regular in English spelling and conform to a basic **CVC** or **CVCC** structure. Finally, we added and removed some words from the final list to ensure good representation of ARPAbet phonemes in the Test Curriculum .

The system is specifically designed to accommodate a changing curriculum of words. Included in the system repository are assets and animations for 35 additional words that were not used for this study. The specific models described in the following section are applicable to a curriculum of any words for which we can compute a phonemization and a word-vector representation.

3.4 Gaussian Process Regression models of children's word pronunciation

In this section we discuss how a computational tutoring agent can construct Gaussian Process Regression (GPR) models of a student's knowledge, and leverage the clean representation of the GPR posterior in an active learning procedure.

Gaussian Processes are a non-parametric method for regression (i.e., function approximation), and have been widely applied in geospatial statistics (where GPR is also known as 'kriging'). They have also been used for educational tasks and other classification contexts in which training data is scarce. Lindsey and Mozer used a GPR-based model for structuring optimal sequencing of a curriculum [22], and Griffith et al. used GPR to model human learning of rules describing functions [11]. Kapoor et al. combined Gaussian processes with active learning for image classification, demonstrating that GPs can be a computationally efficient way to classify data in a large concept space, given an appropriate kernel for computing the covariance between inputs [18].

Gaussian Processes have a number of computational properties that make them well-suited for modeling student vocabulary skills. First, for each point in the test domain they provide Gaussian outputs, defined by a posterior mean and posterior variance. In other words, unlike non-probabilistic discriminatory methods (e.g., Support Vector Machines), Gaussian Processes quantify the uncertainty in their classification. This has a natural semantic interpretation in our models as a measure of the expected pronunciation score for a given word or phoneme, as well as the uncertainty about that estimate, which we leverage in the active learning protocol to select the most appropriate word for the current objective.

Second, GPR models defined over one test domain can be easily extended to another (still finite) test domain, so long as we can compute the covariance between observed data points and the new test points (in this case, words for which we can compute phonemic and/or word-vector representations), without additional training data. In other words, after training, GPR models generalize well to other words "near" the original training data, *even if not explicitly modeled when training occurred.*

Finally, GPR models make efficient use of data compared to other state-of-the-art learning models (e.g., deep neural networks). The GPR model training step takes $O(n^3)$ time, which is often impractical for very large datasets, but is computationally acceptable and provides good performance on tasks where there are few labeled examples [31]. This is often the case with personalized educational models, especially speech-based models.

These properties makes Gaussian Process Regression an intriguing tool for quickly learning models of student knowledge from few data points. In the next section we review the technical details of GPR models in general, and discuss the training and construction of the WORD-GPR and PHONEME-GPR models specifically.

3.4.1 Phoneme and Word GPR models of student pronunciation. Formally, a Gaussian Process is defined by mean and covariance functions, and specifies a distribution over regression functions.

$$f \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x'}) \tag{1}$$

While traditional implementations set $m(x) = 0$, in this case, we use a mean function of $m(x) = 0.5$, reflecting our prior assumption that unobserved words or phonemes are equally likely to be answered correctly or not. After observing data, the mean function contributes nearly nothing to the posterior [29].

One of the key insights that makes Gaussian Process Regression tractable is that while a GP is a distribution over possible functions, we need only define each possible function's values at a finite collection of points, which we refer to here as the *test domain.*

We define two Gaussian Processes: a WORD-GPR model defined over a test domain of **words**, $w \in W$, (where W is the set of words in the Test Curriculum) and a PHONEME-GPR model defined over a test domain of **phonemes**, $p \in P$, (where P is the set of ARPAbet phonemes). We incrementally retrain the GP after each round of the game, and the agent uses the latest posterior means and variances of the model to select the word for the next round.

The training data is a chronologically ordered sequence of observations, $O = \{o_1, o_2 ..., o_t\}$, where each observation, o_t, is a word-score pair, $\{w_i, \{s_{p0}, ...s_{pj}, \}\}$ consisting of a word from the Test Curriculum and a set of scores for the pronunciation of each phoneme in the word, evaluated by SpeechAce (see Sec 3.2).

The GP posterior at time t is a mapping from a (finite) set of points in the test domain to a set of normally distributed random variables, $GP_t : (\{x_i\} \in X) \rightarrow \{N_i^t\}$, with each $N_i^t = \{(\mu_i^t, \sigma_i^t)\}$. Throughout this paper, we refer to μ_i^t and σ_i^t as the posterior mean and posterior variance of word w_i at time t.

For words that we have observed directly, this is straightforward Maximum A Posteriori estimation – we assume that the child's answer is generated by a normal random variable and calculate the most likely mean and variance, given the data we have observed. For words that we have not observed directly, we can still update our posterior estimate: we define a kernel (covariance) function that specifies how "close" or "far" each word is from each other. In this way, we can get more information out of few samples and quickly learn a good model across all words in the vocabulary.

3.5 Phonetic / Semantic Distance Kernel

A GPR model's behavior is largely defined by its covariance function. Hence, to apply GPR to word and phoneme pronunciation, we must define how "close" each pair of words and phonemes are to each other. In this section, we define two covariance metrics, one based primarily on phonemic similarity (for the PHONEME-GPR model) and another for the WORD-GPR model that combines the phonemic similarity metric with a *semantic* similarity metric, based on the pedagogical reality that words are learned, recognized, and recalled as conceptual entities in a semantic graph.

3.5.1 Phoneme-based Covariance. Some pairs of phonemes feel intuitively "more similar" than others. To quantify this intuition, Hixon et al. derived a Weighted Phoneme Substitution Matrix (WPSM) for American English by analyzing multiple acceptable pronunciations for words (e.g., To-MAY-To vs To-MAH-To) [16]. Instances of two phonemes being substituted in alternate pronunciations bring the distance between those two phonemes 'closer' to each other. For the covariance function for the PHONEME-GPR models, we used a normalized version of Hixon's WPSM (Eq. 2).

3.5.2 Word-based Covariance. The covariance function for the WORD-GPR model is a weighted sum of two word-distance metrics: a phonemic distance metric and a semantic distance metric. Intuitively, if we observe that the child can correctly pronounce "CAT" then we believe they are likely to also know how to pronounce "CAR" because they are phonetically similar, and also "DOG" because they are semantically similar.

The phoneme distance between two words is derived by computing the Levenshtein distance between phonemized representations of words (i.e., the minimum-cost combination of additions, deletions, and substitutions necessary to transform one word into another). By convention, addition and deletion costs are fixed at 1. We use the normalized WPSM from the PHONEME-GPR covariance function as the substitution costs for each set of phonemes (Eq. 3)

The semantic distance between two words is the cosine distance between their word-embedding vector representation. We used word-embedding vectors from the pre-trained Common Crawl GloVe vectors [27], included with the SpaCy Python package.

The Levenshtein distances between words are normalized to [0, 1] and combined with a semantic word-distance metric to determine

the final covariance between each pair of words. For this evaluation, we set $\alpha = .33, \beta = .66$ (Eq. 4).

$$k(p_1, p_2) = WPSM(p_1, p_2) \rightarrow [0, 1] \quad (2)$$

$$WLD(w_1, w_2) = Levenshtein(w_1, w_2, WPSM) \quad (3)$$

$$k(w_1, w_2) = \alpha WLD(w_1, w_2) + \beta Cos(GloVe(w_1), GloVe(w_2)) \quad (4)$$

4 EXPERIMENT AND ANALYSIS DESIGN

We recruited 15 children from the local area to participate in a system evaluation interaction in three phases: PRACTICE, EXPERIMENT, and POSTTEST. After arriving at our lab, children were asked to draw a picture of a robot as a 'warm-up' exercise while the parent filled out a survey and consent form. The child was introduced to the robot and the experimenter explained the rules and interface of the game. For this experiment, we used a commercially available social robot, though any ROS-compatible robot platform can be used, as the agent's gameplay behaviors are abstracted through a software interface. After the participant was introduced to the robot, they were invited to play 7 practice rounds of the game with the robot while the experimenter gave advice (not always followed) to speak slowly and clearly and to help children understand the structure and timing of the game. The robot's behavior was fully scripted during the practice round to allow the child opportunities to ring in and practice using the game interface, as well as see the robot ring in and pronounce a word. The PRACTICE phase served as an attempt to help mitigate common issues with analyzing children's speech, such as disfluencies, repetition, or clipping in the recordings. None of the words used in the PRACTICE phase were used in the EXPERIMENT or POSTTEST phases.

After completing the PRACTICE phase, participants were asked if they wanted to practice again or move on to the EXPERIMENT phase. With a few exceptions, participants chose to move on to the game immediately. The experimenter then launched the full system, including the robot controller, the Agent and Student models, and the sensors. Participants played the game with the robot until they had given 20 pronunciation demonstrations, or unless they wished to stop, in which case their data was not included in the analysis.

After the EXPERIMENT phase, the experimenter conducted a short qualitative survey with participants as a break. Once the survey was completed, participants moved on to the POSTTEST phase, in which then played a version of the game without an opposing agent player. Using the same tablet interface, but without the robot across from them, children were presented with each word in the Test Curriculum in sequence and allowed to ring-in and pronounce the word, as before. The POSTTEST phase was intended to emulate the interaction style of 'gamified' workbook-style software, with animations, scores, and sound effects, but no agent-based interaction. After the POSTTEST was completed or the child indicated they did not want to play any more, the interaction ended and the child was given a choice of several stickers as a thank-you for participating.

4.1 Ground Truth Human ratings

We used the audio recordings from the POSTTEST phase to derive a "ground truth" of pronunciation quality from human raters using Amazon Mechanical Turk (MTurk). Each of the recordings was posted as a Human Intelligence Task (HIT) on MTurk. Each posted

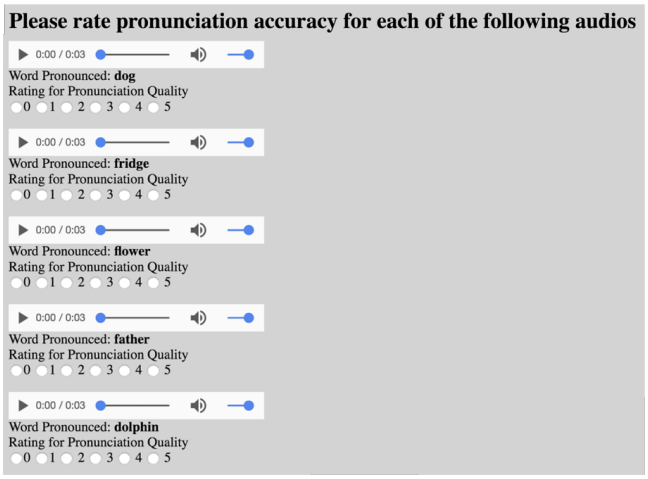

Figure 3: Example HIT interface for human raters

HIT contained five different recordings and only one worker was allowed to submit each HIT. No MTurk workers could rate the same audios more than once. To ensure rating quality, we set the following qualifications: (1) a qualified worker must be geographically located in either the United States or Canada (2) a qualified worker must have a HIT approval rate higher than 90% (3) a qualified worker must have completed at least 100 HITs in the past. Only workers who satisfied all three requirements were considered eligible pronunciation evaluators. To reduce inter-coder variability, qualified workers were required to listen to five 'reference' audios (i.e., exemplars of each class) before they completed each HIT. A screenshot of our rating HIT interface is displayed in Figure 3.

In order to obtain reliable, consistent rating scores for our speech samples, our rating system only finalizes a score for a recording if it receives two ratings of the same score and the absolute difference between this score and its nearest-neighbor score is no larger than 1. The rating system initially sent the full set of collected speech audios to MTurk to receive 3 independent ratings for each audio. After workers completed all HITs, the system evaluated the ratings for all audios, and re-uploaded recordings that did not meet the rating consistency criteria as new HITs to obtain additional ratings. This process repeated until all audios had consistent, final scores.

5 EVALUATION, RESULTS, AND DISCUSSION

We evaluate the classification performance of the WORD-GPR and PHONEME-GPR models by using the final posterior mean (after the EXPERIMENT phase) of each word in the Test Curriculum as a classification probability to predict the pronunciation acceptability labels provided by human raters from POSTTEST phase recordings. The models are trained on SpeechAce scores from the EXPERIMENT phase, however we also provide a side-by-side comparison of how well SpeechAce scores from the POSTTEST phase predict the human labels from the POSTTEST phase. This comparison serves as both an evaluation of the SpeechAce software itself and a rough bound for how well we might reasonably expect the GPR models to perform.

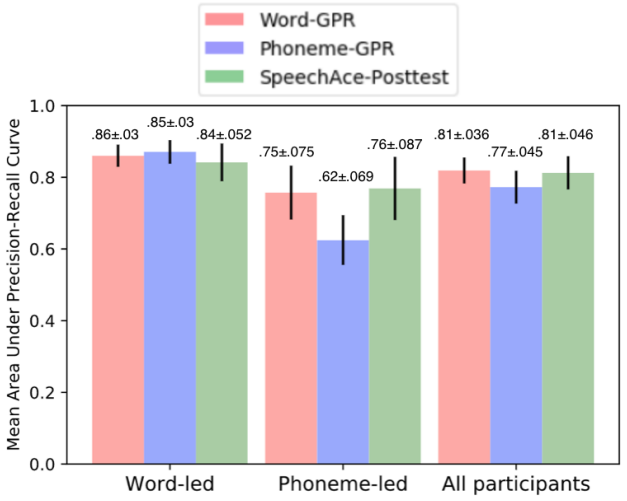

Figure 4: Classification results (total area under the Precision-Recall curve) of Word- and Phoneme- GPR models trained on data from the Experiment phase. SpeechAce results were derived directly from Posttest data.

To evaluate the GPR models, we conducted an area-under-the-curve (AUC) analysis, a common evaluation technique for probabilistic binary classifiers that balances the rate of True and False Positives across a variety of classification thresholds.

For AUC analysis, the true class labels must be binary. Therefore, we binarized the expert ratings such that pronunciations with a rating of 3, 4, or 5 indicate an **acceptable** pronunciation, while a rating of 0, 1, or 2 indicated an **unacceptable** pronunciation. Under this binarization, the final ground truth test set had 396 **acceptable** (positive) examples and 148 **unacceptable** (negative) examples. This type of class imbalance can skew the more common Receiver Operating Characteristic (ROC) AUC; our AUC results are therefore derived from the area under the Precision-Recall Curve, a more balanced metric for datasets with significant class imbalances [7].

Figure 4 shows the AUC scores of both WORD-GPR and PHONEME-GPR models, in both the Phoneme-Led and Word-Led conditions, alongside the AUC of a classifier derived directly from SpeechAce. For each audio recorded during the POSTTEST phase, we use the word-score received from SpeechAce to predict the ground truth rating label of the same audio. All analyses were implemented using *scikit-learn*, a freely-available Python library with implementations of many common data science and machine-learning algorithms [26]. Error bars represent standard error of the mean (SEM).

5.1 Experiment-trained GPR Models accurately predict Children's Posttest performance

The results show that SpeechAce can be used to classify children's pronunciation as acceptable or not (avg. AUC = .81). More interestingly, the GPR models trained on data from the EXPERIMENT phase yield nearly as good classification performance, despite being trained on speech samples from a different experimental phase and with training data less than half the size of the full Test Curriculum (20 vs. 45). These results suggest that GPR can be an effective and efficient way to quickly assess children's productive vocabulary for educational modeling. The WORD-GPR models overall perform better (avg. AUC = .81) than the PHONEME-GPR models (avg. AUC = .77) and all models trained in the Word-Led condition perform better than models trained in the Phoneme-Led condition. This computational result supports the pedagogical guidance that informed our WORD-GPR covariance kernel: that word pronunciations are learned and practiced in a semantic context, that children's word pronunciation mastery is not a linear function of phoneme mastery, and that learning to read is a more complex process than simply identifying and pronouncing the correct phonemes.

5.2 Limitations of GPR and Future Work

GPs assume the training data comes from a static function and thus struggle to model changes in a student's knowledge that may occur during the interaction. If a child pronounces a word incorrectly, the active learning protocol will correctly select that word for a future demonstration by the robot. After the demonstration, however, the posterior variance of that word remains low, and thus is not likely to be selected by the active learning protocol as a word to let the student try again. While previous work has shown that these sequences of pedagogical behavior patterns (prompt for a word, observe incorrect answer, demonstrate correct answer, prompt word again) can be learned from data [20], the nature of the GP Model (variance does not substantially increase) and the current Active Learning protocol (words with low variance are less likely) interact such that we see few examples of this behavior.

We therefore recommend GPs as a method for quickly assessing student knowledge when it may be too expensive, time-consuming, or impractical to observe/evaluate the student's knowledge across a large curriculum. Other techniques, such as Bayesian Knowledge Tracing (BKT), are better suited to assessing *change* in student knowledge, and we are intrigued by the ways different models could complement each other. For instance, GPR could be used to quickly assess a child's pronunciation ability, and the final model could be used as a personalized prior to a BKT model better suited to representing change in knowledge throughout an interaction.

6 CONCLUSION

In this paper we present an integrated system for efficiently and autonomously assessing children's word pronunciation in the context of an educational game with a social robot. We also presented an open-source game environment (the "Word Racer" game) and an accompanying open-source architecture for integrating and evaluating autonomous tutoring agents and student modeling algorithms. We have evaluated this system with an agent model that supports efficient assessment of the student's knowledge via active learning and phoneme- and word-based GPR models of student vocabulary learned from children's speech, and have shown that the system can effectively assess children's pronunciation ability from speech captured during autonomous, interactive gameplay.

ACKNOWLEDGMENTS

The authors would like offer thanks to Kim Hui and Aya Fukuda for invaluable design and artistic help, to the SpeechAce team in Seattle for providing us a fast, accurate, and novel pronunciation assessment tool, to Huawei Co. for their generous support, and to the anonymous reviewers for their valuable comments and helpful suggestions. This work was supported by an NSF Graduate Research Fellowship and Grant IIS-1734443.

REFERENCES

[1] Ryan Shaun Baker, Albert T Corbett, Kenneth R Koedinger, and Angela Z Wagner. 2004. Off-task behavior in the cognitive tutor classroom: when students game the system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 383–390.

[2] Paul Baxter, Emily Ashurst, Robin Read, James Kennedy, and Tony Belpaeme. 2017. Robot education peers in a situated primary school study: Personalisation promotes child learning. *PLoS One* 12, 5 (2017), e0178126.

[3] Mary E Beckman, Andrew R Plummer, Benjamin Munson, and Patrick F Reidy. 2017. Methods for eliciting, annotating, and analyzing databases for child speech development. *Computer Speech & Language* 45 (2017), 278–299.

[4] Andrew Biemiller and Naomi Slonim. 2001. Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of educational psychology* 93, 3 (2001), 498–520.

[5] Ginevra Castellano, Iolanda Leite, André Pereira, Carlos Martinho, Ana Paiva, and Peter W Mcowan. 2013. Multimodal affect modeling and recognition for empathic robot companions. *International Journal of Humanoid Robotics* 10, 01 (2013), 1350010.

[6] Christopher Crick, Graylin Jay, Sarah Osentoski, Benjamin Pitzer, and Odest Chadwicke Jenkins. 2017. Rosbridge: Ros for non-ros users. In *Robotics Research*. Springer, 493–504.

[7] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 233–240.

[8] Margaret Fish and Brenda Pinkerman. 2003. Language skills in low-SES rural Appalachian children: Normative development and individual differences, infancy to preschool. *Journal of Applied Developmental Psychology* 23, 5 (2003), 539–565.

[9] Goren Gordon and Cynthia Breazeal. 2015. Bayesian Active Learning-based Robot Tutor for Children?s Word-Reading Skills. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI-15)*.

[10] Goren Gordon, Samuel Spaulding, Jacqueline Kory Westlund, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. 2016. Affective Personalization of a Social Robot Tutor for Children's Second Language Skills. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[11] Thomas L Griffiths, Chris Lucas, Joseph Williams, and Michael L Kalish. 2009. Modeling human function learning with Gaussian processes. In *Advances in neural information processing systems*. 553–560.

[12] Andreas Hagen, Bryan Pellom, and Ronald Cole. 2003. Children's speech recognition with application to interactive books and tutors. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. IEEE, 186–191.

[13] Andreas Hagen, Bryan Pellom, Sarel Van Vuuren, and Ronald Cole. 2004. Advances in children's speech recognition within an interactive literacy tutor. In *Proceedings of HLT-NAACL 2004: Short Papers*. Association for Computational Linguistics, 25–28.

[14] Betty Hart and Todd R Risley. 1995. *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

[15] Scott Heath, Gautier Durantin, Marie Boden, Kristyn Hensby, Jonathon Taufatofua, Ola Olsson, Jason Weigel, Paul Pounds, and Janet Wiles. 2017. spatiotemporal aspects of engagement during Dialogic storytelling child–robot interaction. *Frontiers in Robotics and AI* 4 (2017), 27.

[16] Ben Hixon, Eric Schneider, and Susan L Epstein. 2011. Phonemic similarity metrics to compare pronunciation methods. In *Twelfth Annual Conference of the International Speech Communication Association*.

[17] SpeechAce Inc. 2017. SpeechAce. (November 2017). [Online; Accessed November 1, 2015].

[18] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. 2010. Gaussian processes for object categorization. *International journal of computer vision* 88, 2 (2010), 169–188.

[19] James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*. ACM, New York, NY, USA, 82–90. https://doi.org/10.1145/2909824.3020229

[20] W Bradley Knox, Samuel Spaulding, and Cynthia Breazeal. 2016. Learning from the wizard: Programming social interaction through teleoperated demonstrations. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1309–1310.

[21] Manoj Kumar, Daniel Bone, Kelly McWilliams, Shanna Williams, Thomas D Lyon, and Shrikanth Narayanan. 2017. Multi-scale Context Adaptation for Improving Child Automatic Speech Recognition in Child-Adult Spoken Interactions. *Proc. Interspeech 2017* (2017), 2730–2734.

[22] Robert V Lindsey, Michael C Mozer, William J Huggins, and Harold Pashler. 2013. Optimizing instructional policies. In *Advances in Neural Information Processing Systems*. 2778–2786.

[23] National Center for Educational Statistics. 2017. The Condition of Education 2015 (NCES 2015-144). National Center for Education Statistics. Washington, DC. (Jan 2017). https://nces.ed.gov/pubs2015/2015144.pdf

[24] Mariela M Páez, Patton O Tabors, and Lisa M López. 2007. Dual language and literacy development of Spanish-speaking preschool children. *Journal of applied developmental psychology* 28, 2 (2007), 85–102.

[25] Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal. 2017. Telling Stories to Robots: The Effect of Backchanneling on a Child's Storytelling. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 100–108.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[27] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[28] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source Robot Operating System. In *ICRA workshop on open source software*. 5.

[29] Carl Edward Rasmussen and Christopher K. I. Williams. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

[30] Najmeh Sadoughi, André Pereira, Rishub Jain, Iolanda Leite, and Jill Fain Lehman. 2017. Creating Prosodic Synchrony for a Robot Co-player in a Speech-controlled Game for Children. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 91–99.

[31] Yirong Shen, Matthias Seeger, and Andrew Y Ng. 2006. Fast gaussian process regression using kd-trees. In *Advances in neural information processing systems*. 1225–1232.

[32] Catherine E Snow, Michelle V Porche, Patton O Tabors, and Stephanie Ross Harris. 2007. Is literacy enough? Pathways to academic success for adolescents. (2007).

[33] Samuel Spaulding, Goren Gordon, and Cynthia Breazeal. 2016. Affect-aware student models for robot tutors. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 864–872.

[34] Adriana Tapus, Maja Mataric, Brian Scassellatti, et al. 2007. The grand challenges in socially assistive robotics. *IEEE Robotics and Automation Magazine* 14, 1 (2007).

[35] Joseph Tepperman, Jorge F Silva, Abe Kazemzadeh, Hong You, Sungbok Lee, Abeer Alwan, and Shrikanth Narayanan. Pronunciation verification of children^{2}s speech for automatic literacy assessment.

[36] Robert L Weide. 1998. The CMU pronouncing dictionary. *URL: http://www. speech. cs. cmu. edu/cgibin/cmudict* (1998).

[37] Beverly Park Woolf. 2010. *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.

[38] Gary Yeung, Amber Afshan, Kaan Ege Ozgun, Kantapon Kaewtip, Steven M Lulich, and Abeer Alwan. Predicting Clinical Evaluations of Children's Speech with Limited Data Using Exemplar Word Template References.